Dual-view Curricular Optimal Transport for Cross-lingual Cross-modal Retrieval

Yabing Wang, Shuhui Wang, Hao Luo, Jianfeng Dong, Fan Wang, Meng Han, Xun Wang, and Meng Wang

Abstract—Current research on cross-modal retrieval is mostly English-oriented, as the availability of a large number of English-oriented human-labeled vision-language corpora. In order to break the limit of non-English labeled data, cross-lingual cross-modal retrieval (CCR) has attracted increasing attention. Most CCR methods construct pseudo-parallel vision-language corpora via Machine Translation (MT) to achieve cross-lingual transfer. However, the translated sentences from MT are generally imperfect in describing the corresponding visual contents. Improperly assuming the pseudo-parallel data are correctly correlated will make the networks overfit to the noisy correspondence. Therefore, we propose Dual-view Curricular Optimal Transport (DCOT) to learn with noisy correspondence in CCR. In particular, we quantify the confidence of the sample pair correlation with optimal transport theory from both the cross-lingual and cross-modal views, and design dual-view curriculum learning to dynamically model the transportation costs according to the learning stage of the two views. Extensive experiments are conducted on two multilingual image-text datasets and one video-text dataset, and the results demonstrate the effectiveness and robustness of the proposed method. Besides, our proposed method also shows a good expansibility to cross-lingual image-text baselines and a decent generalization on out-of-domain data.

Index Terms—Cross-modal retrieval, Noise correspondence learning, Cross-lingual transfer, Optimal transport, Machine translation.

I. INTRODUCTION

CROSS-LINGUAL Cross-modal Retrieval (CCR) retrieves the visual contents (*i.e.,* videos or images) which are semantically relevant based on target-language (*e.g.,* non-English) queries T, but can only be trained on the manually annotated pairs of visual contents V and source language (*e.g.,* English) captions S. It aims to alleviate the problem of the existence of large-scale multilingual vision-language corpora and the limited development of non-English languages in the field of cross-modal retrieval [1]–[10].

The key of CCR is how to achieve effective cross-lingual transfer to facilitate alignment between visual and target-language features. Recently, a series of breakthroughs have been proposed [11]–[19]. Instead of relying on the parallel corpus for direct visual-target language alignment, some

Y. Wang, J. Dong and X. Wang are with the College of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310035, China.
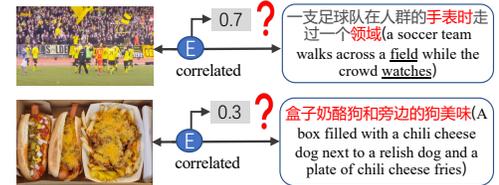S. Wang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.
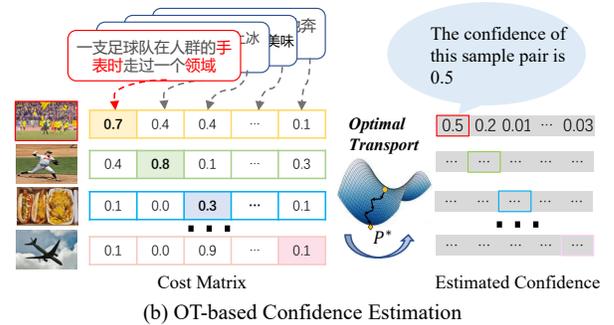H. Luo and F. Wang are with Alibaba Group, Hangzhou 310052, China.
M. Han is with the Binjiang Institute of Zhejiang University, Hangzhou 310027, China.
M. Wang is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China.

Fig. 1. An example of the P2P-based method and our proposed OT-based confidence estimation method. The red words represent the incorrectly translated ones.

works [16], [17] utilize source language as the focal point to build a bridge between visual content and target language. However, they fail to break the semantic gap between the visual and target language and the parallel corpus is still costly to collect. With the popularity of Machine Translation (MT), a natural solution [14], [19] is to generate pseudo visual and target language pairs by MT and directly establish their correspondences. In specific, [14], [19] pre-train the model with a large number of pairs of visual data and translated target-language captions (V+T). However, they still rely on large-scale vision-language datasets (*e.g.,* CC3M [20] and its translation) and ignore the noise from translation. As shown in Fig. 1 (a), even with the most powerful off-the-shelf MT tools, the translated target-language captions still contain various noises, such as spelling errors, grammar errors, and even distorted overall meaning.

Due to the noise introduced during the translation process, the imperfect target-language captions cannot accurately describe the corresponding visual contents (*i.e.,* noise correspondence problem). In this case, if we persist in promoting the alignment between the visual and target-language features in a common space, the model will overfit to the wrong supervision and result in degraded performance. A recent method called Noise-Robust Cross-lingual Cross-modal Retrieval (NRCCR) [18] employs multi-view distillation to gener-

ate soft pseudo-targets as direct supervision for target-language learning, and achieves comparable results to methods of using extra pre-training data. Considering the ubiquitousness of noisy correspondence in various cross-domain matching tasks, there have been consistent endeavors to alleviate its adversarial influence. A typical solution is to use Cosine or Euclidean distance as the point-to-point (P2P) correspondence of sample pairs [2], [21]. However, similar with NRCCR, these methods (Fig. 1 (a)) ignore the instance relation in context points in two sets of data, while the context information between sample pairs is crucial for reliable correspondence. Some works [22], [23] utilize the Gaussian Mixture Model (GMM) to divide the data into clean and noisy partitions. Although GMM considers the distribution relation between samples, the assumption of mixture of Gaussian may not capture complex and diverse patterns in real-world data (*e.g.,* long-tail distribution). Moreover, these methods only address the matching problem between two sets of instances ($V \leftrightarrow T$ or $S \leftrightarrow T$), which is not well-suited for CCR with multiple domains ($V \leftrightarrow T \leftrightarrow S$).

To tackle the aforementioned limitations, this paper proposes a CCR-specific noise-robust method called Dual-view Curricular Optimal Transport (DCOT). To obtain reliable correspondence, we formulate the noisy correspondence learning as an optimal transport (OT) problem. Instead of finding correspondences between individual points of two sets, our method aims to find an optimal matching between the two sets (Fig. 1 (b)). We interpret the optimal matching as a confidence measure for the correct matching between sample pairs, which allows us to evaluate the reliability of the correlation score between sample pairs. Considering the presence of multiple domains in CCR, we incorporate both cross-lingual and cross-modal views and use OT from both views to quantify the confidence of each correlated sample pair. Through theoretical and empirical analysis, we found that the model fitting is undertaken quickly on cross-lingual view in the early stage, and gradually transferred to cross-modal view in the later stage. Accordingly, we design a dual-view curriculum learning process, which constructs the transportation costs and determines the weights of both views dynamically with a curriculum schedule based on the learning status of the two views at each time-step during training. Our method is more flexible to different types of noise in CCR, as we do not make any assumptions about the underlying data distribution. Our contributions can be summarized as follows:

- To take into account the instance relation in context, we formulate the noisy correspondence learning in CCR as an optimal transport problem.
- The proposed DCOT method dynamically models the transportation costs according to the learning state of two views, *i.e.,* cross-lingual and cross-modal views, to avoid overfitting to the noisy sample pairs.
- Extensive experiments on three image-text and video-text cross-modal retrieval benchmarks across different languages demonstrate the effectiveness and robustness of our method.

II. RELATED WORKS

A. Cross-lingual Cross-modal Retrieval

Cross-lingual transfer learning has become a crucial mechanism to battle the unavailability of annotated low-resource languages. Recently, some works [11]–[18] try to apply cross-lingual transfer learning to cross-modal retrieval tasks to alleviate the problem of data scarcity and achieved remarkable progress. Under the CCR setting, the model trained on manually annotated pairs of vision and source language is adapted for evaluations in different target languages. Prior works [11], [13] aligning different languages into a common space with non-contextualized multilingual word embeddings (MUSE [24] and BIVEC [25]) and pre-trained sentence encoders(mUSE [26] and LASER [24]), respectively. The major study can be divided into three groups based on how the alignment of visual-target language is achieved: 1) rely on parallel corpus, 2) collect multilingual subtitles from the web, and 3) resort to MT.

To be specific, methods that rely on parallel corpus [16], [17] regard English as the focal point to build a bridge between visual and target languages. For example, the code-switched training [17] enforces the explicit alignment between images and non-English languages. However, these methods indirectly align visual and non-English languages centered on English, and the process of collecting parallel corpus is also costly and time-consuming. Huang *et al.* [15] crawl and collect the multilingual subtitles from YouTube, and extend the HowTo100M [27] to the multilingual version. Multilingual Vision-Language data corpus with MT [14], [18], [19] have recently emerged as an alternative. An MT-augmented cross-lingual cross-modal pretraining framework [14] is proposed, which pivots primarily on images and complementarily on English for multilingual multi-modal representation learning. Further, a noise-robust learning framework [18] is proposed to deal with the noise in MT results. However, they do not explicitly evaluate the confidence of the samples but filter the noise in the target language implicitly by introducing a cross-attention module. This may lead to a suboptimal solution for solving the model overfitting in the presence of noisy labels.

B. Learning with Noisy Correspondence

The issue of noisy labels has been well studied in the visual classification task [28]–[30]. For multi-modal learning, a new paradigm [22] is developed, which considers the alignment errors in paired data instead of the errors in category annotations in the classification task. Recently, some research focused on noise correspondence, such as [**?**], [2], [22], [31] in cross-modal retrieval task, and [18] in cross-lingual cross-modal retrieval task. Among them, some works resort to robust architecture design [2], [18], [21], [32]. For example, [18] employs multi-view self-distillation to generate soft pseudo-targets to provide direct supervision for noise-robust target-language representation learning. Other works attempted to evaluate the confidence of sample pairs and design noise-robust losses [22], [23], [31]. For example, in [22], [23], the annotation confidence among the samples is evaluated by introducing the GMM and the soft margin of
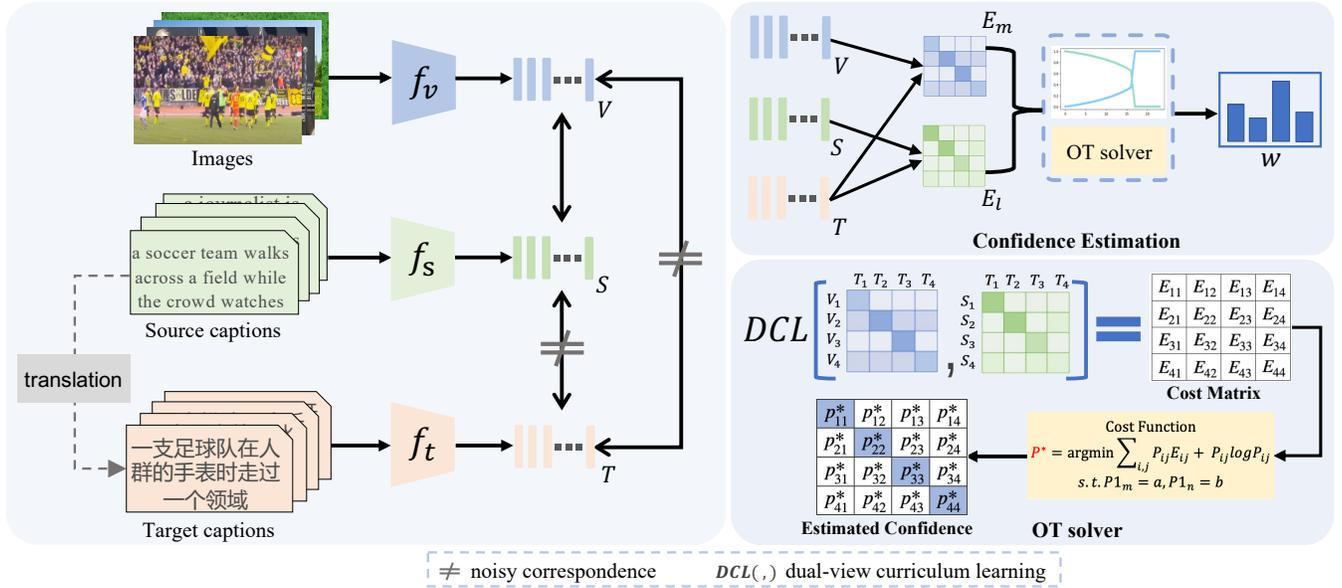
Fig. 2. **Illustration of our proposed DCOT method for CCR.** The images, source-language captions and target-language captions are first encoded to representations. Then the confidence of the correctly correlated for (V, T) pairs is estimated by OT-based confidence estimation from both the cross-lingual view and cross-modal views. For confidence estimation, we design a dual-view curriculum learning, by which the transportation costs modeling is tied to the learning state of the two views. Finally, DCOT adjusts the contribution of each pair dynamically according to the estimated confidence during training.

sample pairs adjusted in the triplet loss. Different from these methods, our method is more generalized and does not rely on any prior information on the input data distribution. It can flexibly combine the cross-lingual and the cross-modal view to estimate the confidence of the correlated sample pairs.

III. METHOD

A. Preliminaries

Let $\mathcal{D} = \{(V_i, S_i)\}_{i=1}^{N}$ be a dataset of annotated paired images/videos and source-language captions with data size N. As the access to human-labeled vision and target language sample pairs during training is unavailable, some external tools can be utilized, *e.g.,*, MT or parallel corpus. Following [18], we extend the training data $\mathcal{D}$ to $\hat{\mathcal{D}} = \{(V_i, S_i, T_i)\}_{i=1}^{N}$ with MT, where T_i is the translated target-language caption corresponding to S_i. We define $f_v(.)$, $f_s(.)$ and $f_t(.)$ as the image/video, source-language and target-language encoder, respectively, and denote the embedded fixed-dimensional vectors of an image/video as V_i, a source-language caption as S_i, and a target-language caption as T_i. Note that we use the human-input target-language sentences as queries for retrieval during inference. The framework is illustrated in Fig. 2. In what follows, we will first introduce our proposed OT-based confidence estimation (Sec. III-B), then describe the dual-view optimal transportation costs modeling strategy (Sec. III-C), and finally introduce the noise-aware alignment objective (Sec. III-D).

B. OT-based Confidence Estimation

Supposing we have a mini-batch with M image/video-caption pairs, denotes $\overline{\mathcal{D}} = (V, S, T)$. Current methods usually utilize the Cosine distance or Euclidean distance to calculate the point-to-point correspondence between individual (V_i, T_i)

pair without confidence estimation, which can be expressed as:

$$E = dist(f_v(V), f_t(T)) \in \mathbb{R}^{M \times M} \quad (1)$$

where $dist(\cdot)$ is the distance function, *e.g.,*, cosine distance, and E_{ij} denotes the correlation score between the i-th visual feature and j-th translated caption feature.

As illustrated in Fig. 3 (a), the pairwise correlation score of (V_i, T_i) is computed individually without context. Since the network parameters change dynamically during the training process, the Cosine distance (or Euclidean distance) of the same feature pair at different time steps turns out to be different. Such a point-to-point correspondence calculation scheme fails to consider the semantic context, and thus it can hardly provide reliable supervision signals, especially under the situation of high noise in the translated sentences. Therefore, to alleviate the noise effect, a better noisy correspondence learning strategy is to consider the mutual relation between features in a data batch to seek for a more reliable contextual correspondence solution.

To address this issue, we propose a solution by formulating noisy correspondence learning as an OT problem and finding the optimal match between data points from two sets. This enables us to capture the correspondence between the sets from a contextual perspective (Fig. 3 (b)), and to estimate the confidence of the correlated sample pairs. Specifically, when the transportation cost between a sample pair is relatively small, their matching degree is higher, indicating that they are highly correlated. Therefore, we can treat the similarity score of the optimal match as a confidence measure for the correct matching between sample pairs. If the similarity score of the optimal match is high, we can consider the matching between sample pairs to be reliable.
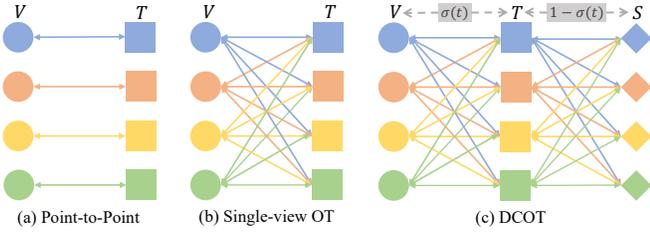
(a) Point-to-Point   (b) Single-view OT   (c) DCOT

Fig. 3. An example of different correspondence calculation methods. The Point-to-Point method (*e.g.,* Cosine distance) does not take into account the contextual relationship between pairs of data. The Single-view OT method estimates the confidence based only on the cross-modal view. In contrast, our proposed method, DCOT, estimates confidence collaboratively from two views.
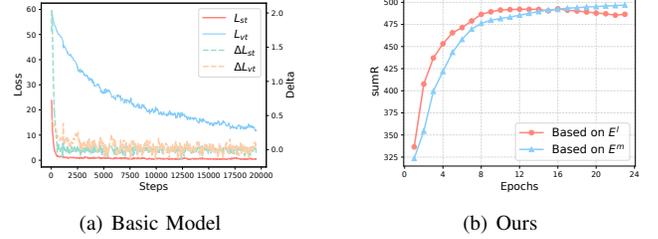

(a) Basic Model   (b) Ours

Fig. 4. (a) We train a baseline model without noise-robust learning (use Eq. (12)) and plot the losses and loss differences Δ incurred by cross-lingual alignment (red and green) and cross-modal alignment (blue and orange) during training. (b) Performance curves during the training on Multi30K. Confidence estimation based on E^l (red) vs. confidence estimation based on E^m (blue).

Optimal transport problem. The Optimal Transport aims to search the most efficient transport plan of transforming one mass distribution to another whilst minimizing the cost. Specifically, given two discrete point sets, $X = \{x_i\}_{i=1}^n$ and $Y = \{y_j\}_{j=1}^m$, $x_i, y_i \in \mathbb{R}^d$. The amount of mass on these points is given by a and b, defined on probability space $X, Y \in \Omega$, respectively.

$$a = \sum_{i=1}^n p_i^x \delta(x_i), b = \sum_{j=1}^m p_j^y \delta(y_j) \qquad (2)$$

where $\delta(\cdot)$ denotes the Dirac function, p_i^x and p_j^y are the probability mass to the i-th and j-th sample, belonging to the probability simplex, *i.e.,* , $\sum_{i=1}^n p_i^x = \sum_{j=1}^m p_j^y = 1$. The unit transportation cost from point x_i to y_j is denoted by C_{ij}. Under such a setting, we aim to search the optimal transport plan P^* to transport the mass in probability measure a to b with the minimum costs by solving the following problem:

$$P^* = \underset{P \in \mathbb{R}^{n \times m}}{\arg\min} \sum_{i,j} P_{ij} C_{ij}$$
$$\text{s.t. } P\mathbb{1}_m = a, \ P^T \mathbb{1}_n = b \qquad (3)$$

where P is the transport plan containing all non-negative $n \times m$ elements with row and column sums to a and b, respectively.

We define E as the transportation costs denoting transporting one unit of translated caption T_i to image/video V_i. Besides, we add an entropic regularization to control the smoothness of the transport plan following [33]. Our goal is to maximize the total correlation to get an optimal estimated confidence on this batch. The corresponding optimization problem can be formulated as follows:

$$\min_{P \in \mathbb{R}^{M \times M}} \sum_{i,j} P_{ij} E_{ij} + \frac{1}{\lambda_{reg}} P_{ij} log P_{ij}$$
$$\text{s.t. } P\mathbb{1}_M = \frac{1}{M}\mathbb{1}_M, \ P^T \mathbb{1}_M = \frac{1}{M}\mathbb{1}_M \qquad (4)$$

where $\lambda_{reg} > 0$ is the regularization parameter, a larger λ_{reg} leads to "softer" distribution for P, and vice versa. Note that the constraint condition ensures that the solution satisfies that all instances in the batch are equally important and should be matched with equal probabilities. The optimal estimated confidence $P^* \in \mathbb{R}^{M \times M}$ of Eq. (4) is:

$$P^* = diag(\mu)Kdiag(v)$$
$$K = \exp(-\lambda_{reg}E) \qquad (5)$$

where u and v are some non-negative vectors, solved with Sinkhorn's fixed point iteration:

$$u^{(t+1)} = \frac{\mathbb{1}_M}{Kv^{(t)}}, \ v^{(t+1)} = \frac{\mathbb{1}_M}{K^T u^{(t+1)}} \qquad (6)$$

Finally, we take the diagonal element w_i of P^* as the confidence of each (V_i, T_i) pair:

$$w = diag(P^*) \qquad (7)$$

C. Dual-view Transportation Cost

As we know, transportation cost is a crucial factor in computing the optimal transport plan. In this section, we introduce a dual-view curriculum learning approach to dynamically model the transportation costs based on the learning state of two views.

Specifically, in CCR, given a (V_i, S_i, T_i) triplet, (V_i, S_i) represents the ground-truth corresponded pair. The correspondence between (V_i, T_i) pair can also be inferred from the correspondence between (S_i, T_i) pair. To obtain more accurate transportation costs, we calculate them from two views, namely the cross-lingual view $(S \leftrightarrow T)$ and the cross-modal view $(V \leftrightarrow T)$, respectively:

$$E^l = dist(f_t(S), f_t(T)) \in \mathbb{R}^{M \times M}$$
$$E^m = dist(f_v(V), f_t(T)) \in \mathbb{R}^{M \times M} \qquad (8)$$

As shown in Fig. 4 (a), the cross-lingual gap is much smaller than the cross-modal one, and the convergence rate of networks in the cross-lingual alignment is faster than that in cross-modal alignment during training. Therefore, confidence estimation based on E^l is more accurate at the beginning of training. Based on this empirical finding, the transportation costs from the cross-lingual view should play a dominant role in transportation cost modeling at the preliminary stage. Besides, considering the impact of memorization effect [34], deep networks tend to first fit the clean sample pairs during an early learning stage before eventually memorizing the wrong sample pairs. Therefore, the networks will gradually memorize noisy (S, T) pairs after quickly learning cross-lingual alignment on clean data pairs. This would affect the accuracy of the confidence estimation based on E^l. On the other hand, the accuracy of the confidence estimation based on E^m increases as the cross-modal alignment improves progressively in the learning process. Therefore, the networks should gradually emphasize the transportation costs from the cross-modal view

in the later stage. To validate this idea, we conducted experiments to explore the influence of the transportation costs of different views on the performance. As Fig. 4 (b) makes clear, the results confirm that the confidence estimation based on E^l achieves better results at the early stage, while that based on E^m tends to perform better at later stage. This is consistent with our assumptions.

Based on this observation, we propose a dual-view curriculum learning strategy to dynamically model the transportation costs from two views collaboratively, as illustrated in Fig. 3 (c). This strategy provides an essential dynamic curriculum where the optimization of transport costs is naturally determined by the learning state of two views. The strategy is formulated as follows:

$$E = \sigma(t)E^m + (1 - \sigma(t))E^l \quad (9)$$

where $\sigma(t) \in [0, 1]$ represents the importance of E^m at time step t. At the beginning of training, the transportation costs would focus on E^l, and then gradually decrease its weight until the transportation costs are dominant by E^m:

$$\sigma(t) = \mathbb{1}(t \leq \tau)h(t \cdot \tau) + \mathbb{1}(t > \tau) \quad (10)$$

where $\tau < 1$ is an empirically-set hyper-parameter that controls the extent of E^l. The function $h(\cdot)$ is a non-linear curriculum to adjust the importance of each view, which can be formulated as:

$$h(z) = \gamma \cdot \frac{z}{2-z} \quad (11)$$

where γ is a hyper-parameter to control the magnitude of change. The curriculum schedule of $\sigma(t)$ ensures that the importance of E^m gradually increases, and equals to 1 when $t > \tau$, which means only E^m is used when $t > \tau$.

Overall, the complete procedure of confidence estimation is presented in Algorithm 1.

D. Noise-aware Alignment Objective

To promote the alignment of cross-lingual and cross-modal, we introduce pairwise alignment loss for given $\{(V_i, S_i, T_i)\}_{i=1}^{M}$ pairs in a mini-batch:

$$\mathcal{L}_{vs} = \sum_{i=1}^{M} \mathcal{L}(f_v(V_i), f_t(S_i))$$
$$\mathcal{L}_{vt} = \sum_{i=1}^{M} \mathcal{L}(f_v(V_i), f_t(T_i)) \quad (12)$$
$$\mathcal{L}_{st} = \sum_{i=1}^{M} \mathcal{L}(f_v(S_i), f_t(T_i))$$

where alignment loss function $\mathcal{L}(,)$ can be implemented by any contrastive loss. Here, we use the triplet ranking loss, which is the major loss objective for cross-modal matching tasks. It enforces the similarity score of the matched visual-text pairs to be larger than the similarity score of the unmatched ones by a margin, formulated as:

$$\mathcal{L}(l_1, l_2) = max(0, r + s(l_1, l_2^-) - s(l_1, l_2))$$
$$+ max(0, r + s(l_2, l_1^-) - s(l_2, l_1)) \quad (13)$$

where l_1 and l_2 denote the input feature vectors, r indicates a margin constant and $s(\cdot)$ denotes the similarity function, *e.g.,* cosine similarity, and l_2^- (or l_1^-) denotes a hardest negative pair for l_1 (or l_2) in the mini-batch.

Algorithm 1 Confidence Estimation Algorithm

Input: Noisy training dataset $\mathcal{D} = \{(V_i, S_i, T_i)\}_{i=1}^{N}$, max epochs T_{max}, iteration I_{max}, maximum iteration number of Sinkhorn algorithm max_iter

Output: w is the optimal estimated confidence of each (V_i, S_i, T_i) pair.

1: **for** $t = 0, 1, ..., T_{max}$ **do**
2:     Shuffle training set
3:     **for** $n = 1, ..., I_{max}$ **do**
4:         Fetch mini-batch $\overline{\mathcal{D}}$ from $\mathcal{D}$
5:         Calculate transportation cost matrices E_m and E_l from two views using Eq.(8)
6:         Calculate dual-view transportation costs E using Eq.(9) with dual-view curriculum learning
7:         Initialize $\mu^{(0)}, v^{(0)}$ as one.
8:         **while** $k < max\_iter$ **do**
9:             Calculate $\mu^{(k+1)}, v^{(k+1)}$ using Eq.(6)
10:        **end while**
11:        Calculate optimal estimation confidence P^* using Eq.(12)
12:     **end for**
13:     $w = diag(P^*)$
14: **end for**
15: **return** w

Given the existence of noisy correspondences in the (V, T) pairs, directly aligning them would result in the model memorizing the noisy correspondence, which would severely degrade its generalizability. Hence, we introduce a noise-aware alignment objective, which adaptively adjusts the contribution of sample pairs based on the estimated confidence score w. This objective function penalizes the noisy sample pairs less, allowing the model to focus on the more reliable samples.

$$\hat{\mathcal{L}}_{vt} = \sum_{i=1}^{M} w_i \mathcal{L}(f_v(V_i), f_t(T_i)) \quad (14)$$

In addition, since the main focus of our task is cross-modal retrieval, we aim to address the noisy correspondence problem in (V, T) pairs. To achieve this, we introduce cross-lingual alignment to assist the target-language encoder in learning the correct semantics from the corresponding source-language captions S. This helps to consistently improve the accuracy of confidence estimation based on E^l. However, we also need to prevent the network from overfitting to noisy (S, T) pairs later in the training process. Therefore, we design a function $G(\cdot)$ that dynamically adjusts the weight of the cross-lingual objective function L_{st}, with the value of $G(\cdot)$ gradually decreasing in the later training stages.

$$\hat{\mathcal{L}}_{st} = G(t)\mathcal{L}_{st}$$
$$G(t) = \frac{1}{1 + ke^{(\epsilon \cdot t - \frac{1}{\tau})}} \quad (15)$$

where k and ϵ are hyper-parameters. Finally, our objectiveness can be formulated as the combination of the above three alignment losses:

$$\mathcal{L} = \hat{\mathcal{L}}_{vt} + \hat{\mathcal{L}}_{st} + \lambda \mathcal{L}_{vs} \quad (16)$$

where λ determines the weight on the alignment task of

images/videos and source-language captions.

IV. EXPERIMENTS

A. Experimental Settings

Datasets. We conduct experiments on two public multi-lingual image-text retrieval datasets: Multi30K [35] and Multi-MSCOCO, which are the multi-lingual version of Flickr30K [36] and MSCOCO [37], respectively, and a public multi-lingual video-text retrieval dataset VATEX [38]. Noting that all non-English captions used in our training are produced by MT instead of human annotations, but human annotations are used as queries during inference.

Multi30K is built by extending Flickr30K [36] from English to German, French and Czech. It contains 31,783 images and provides five captions per image in English and German and one caption per image in French and Czech. The human-labeled test data is provided.

Multi-MSCOCO is extended by MSCOCO [37], and we name it Multi-MSCOCO for ease of reference. It contains 123,287 images, and each image has 5 captions. We translate the training set from English into Japanese and Chinese by resorting to MT, and follow the data split as in [14].

VATEX is a large-scale multi-lingual video dataset. Each video has 10 English captions and 10 Chinese captions to describe the video content. Note that we only use human-labeled English captions for training. Following [39], we split the data into 25,991/1,500/1,500 as train/dev/test.

Evaluation metrics. Following [18], for cross-lingual image-text retrieval, we compute the sum of all R@K ($K = 1, 5, 10$) for both image-to-text and text-to-image retrieval, and use it (sumR) for performance comparison. For cross-lingual video-text retrieval, we measure rank-based performance by R@K ($K = 1, 5, 10$) and sumR for both video-to-text and text-to-video retrieval.

Implementation Details. For image encoder, we use the CLIP (ViT-B/32) [40], a pre-trained language-image model, to extract image representations. For video encoder, we adopt 1,024-dimensional I3D [41] video features and use multi-layer perceptron followed by mean-pooling. For text encoder, we use the pre-trained mBERT-base [42], and take the outputs of the [CLS] token from 9-th layer as the sentence representations.

For model training, we utilize an Adam optimizer and a mini-batch size of 128. The initial learning rate is set to $2.5e-5$. We take an adjustment schedule similar to [43]. For some hyper-parameters during training, we set the ϵ, k and λ as 10, 1 and 0.5 respectively. For multi30K, we set the scaling parameters of dual-view curriculum learning τ and γ as 0.1 and 0.2 respectively. For Multi-MSCOCO, we set them as 0.065 and 0.6 respectively. For VATEX, we set them as 0.065 and 0.1 respectively. We use the same similarity calculation method with NRCCL during inference

B. Ablation Studies

We perform ablation studies on Multi30K to demonstrate the effectiveness of our proposed method.

Effectiveness of dual-view collaboration. As shown in Tab. I, the first row reports the performance of the baseline

TABLE I
ABLATION STUDY OF CONFIDENCE ESTIMATION BASED ON DIFFERENT VIEWS ON MULTI30K. THE METHODS WITH THE CHECKMARK (✔) INTRODUCE NOISE-ROBUST LEARNING. THE METRIC IS THE SUM OF RECALLS (SUMR). "EN", "DE", "FR" AND "CS" INDICATE ENGLISH, GERMAN, FRENCH AND CZECH, RESPECTIVELY.

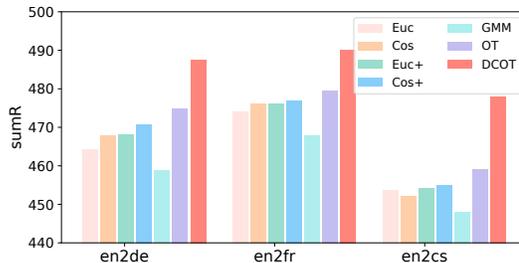| Cross-lingual	Cross-modal	en2de	en2fr	en2cs
✗	✗	476.6	480.7	470.1
✔	✗	481.2	483.3	475.5
✗	✔	482.6	484.0	475.6
✔	✔	**487.4**	**490.2**	**478.1**

Fig. 5. Performance comparison of P2P-based methods, GMM-based method and our proposed OT-based confidence estimation method on Multi30K. "Euc" and "Cos" denote the Euclidean and Cosine distance, respectively. The symbol "+" denotes that a softmax operation is applied after the distance calculation.

method, which is trained only using the loss of Eq. (12). It assumes that all sample pairs are correctly correlated without any noise-robust designs. Compared with the baseline, other methods with noise-robust learning have achieved performance improvement, which suggests that directly promoting the alignment of (V, T) pairs will cause the neural networks to overfit the wrong supervision and degrade the performance. In addition, the performance is significantly improved when we combine the two views, proving the effectiveness and complementarity of our proposed dual-view collaboration.

Effectiveness of OT-based confidence estimation. To validate the effectiveness of the confidence estimation using OT, we compare it with the counterparts using P2P calculation without context (*i.e.,* Euc, Cos, Euc+, and Cos+) and GMM in confidence estimation. As shown in Fig. 5, the OT-based methods achieve significant advantages in all languages. The P2P-based methods suffer from the lack of contextual information, which cannot provide accurate confidence estimation. In contrast, our OT-based algorithm could estimate confidence based on the principle of minimum global costs, by taking the mutual relation between the samples into account. Moreover, the performance of the GMM-based method is severely degraded when the noise distribution deviates from the Gaussian distribution, given that the method has strong constraints on the data distribution. Compared with it, our OT-based method does not require any assumptions about the data distribution, making it more versatile. Additionally, DCOT beats the other methods by a large margin, demonstrating the importance of collaborative effort between the two views, especially in the low-resource languages (*e.g.,* Czech in en2cs). The promising results validate that our proposed confidence estimation algorithm can achieve more accurate confidence
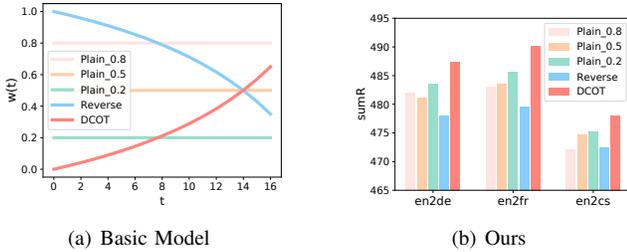
Fig. 6. (a) Examples of various curriculum schedules. We only show the curve in the cross-modal view for ease of representation. The curve in the cross-lingual view corresponds to $1 - w(t)$. (b) Performance comparison of various curriculum schedules on Multi30K.

TABLE II
PERFORMANCE COMPARISON OF CROSS-LINGUAL IMAGE-TEXT RETRIEVAL ON MULTI30K (THE SOURCE LANGUAGE IS ENGLISH AND THE TARGET LANGUAGE IS NON-ENGLISH). THE SCORES ARE THE SUM OF ALL RECALLS (SUMR). THE SYMBOL ASTERISK (*) INDICATES THAT MODEL WAS PRE-TRAINED ON LARGE-SCALE DATASETS, E.G., CC3M AND ITS MT VERSION.

Method	Backbone (#parameters)	en2de	en2fr	en2cs
M^{3}P [17]*	XLMR-large (560M)	351.0	276.0	220.8
UC^2 [14]*	XLMR-base (278M)	449.4	444.0	407.4
CCLM [19]*	XLMR-large (560M)	503.4	490.6	481.6
NRCCR [18]	mBERT (170M)	480.6	482.1	467.1
DCOT(ours)	mBERT (170M)	494.9	495.3	481.8
CCLM+ours*	XLMR-large (560M)	**515.2**	**518.7**	**512.1**

and greatly enhance the robustness of the model to noise.

Influence of curriculum schedule. To perform an in-depth study on dual-view curriculum learning, we conduct experiments with various curriculum schedules shown in Fig. 6 (a) and the results are displayed in Fig. 6 (b). As we can see, the reverse schedule performs the worst. The reason lies in that the collaboration process of two views exactly deviates from the relative accuracy of the confidence estimation based on the two views, leading to a significant performance drop. For plain schedules, the importance of each view remains fixed throughout the training. In contrast, our proposed dynamic schedule adjusts the importance of each view according to their learning state, which has a significant superiority. Thus, designing the appropriate schedule is crucial in improving the accuracy of the confidence estimation.

C. Comparison with State-of-the-Arts

1) Cross-lingual Image-Text Retrieval: For cross-lingual image-text retrieval, we compare four state-of-the-art (SOTA) methods, including M^{3}P [17], UC^2 [14], CCLM [19], and NRCCR [18]. Among them, M^{3}P, UC^2, and CCLM are all pre-trained on the large-scale vision-language corpus, while NRCCR is the robust learning method against noisy correspondence. For a fair comparison, we compare DCOT to CCLM with the dual-stream structure, as the dual-stream models are more suitable for large-scale retrieval.

Comparisons on Multi30K. Tab. II summarizes the performance comparison on Multi30K. Without pre-training and using a more lightweight backbone, DCOT outperforms the large-scale pre-trained model M^{3}P and UC^2 by a large margin,

TABLE III
PERFORMANCE COMPARISON OF CROSS-LINGUAL IMAGE-TEXT RETRIEVAL ON MULTI-MSCOCO. "ZH" AND "JA" INDICATE THE CHINESE AND JAPANESE, RESPECTIVELY.

Method	Backbone (#parameters)	en2zh	en2ja
M^{3}P [17]*	XLMR-large (560M)	322.8	336.0
UC^2 [14]*	XLMR-base (278M)	492.0	430.2
CCLM [19]*	XLMR-large (560M)	511.2	496.4
NRCCR [18]	mBERT (170M)	512.4	507.0
DCOT(ours)	mBERT (170M)	521.5	515.3
CCLM+ours*	XLMR-large (560M)	**535.6**	**536.2**

and achieves comparable performance to CCLM. Moreover, the sumR scores of DCOT is 2.2%, 2.7% and 3.1% higher than noise-robust learning baseline NRCCR on three languages, respectively. As NRCCR obtains pseudo supervision signals by calculating point-to-point correspondence of (V, T) pairs, it does not take the mutual relation between features into account. By contrast, our DCOT models the confidence estimation from a contextual perspective. The results demonstrate that our proposed contextual modeling is more beneficial for noisy correspondence learning.

Recall that our proposed dual-view curricular optimal transport is orthogonal to cross-lingual image-text similarity learning. In this experiment, we evaluate its expansibility to cross-lingual image-text baselines. Specifically, we employ our dual-view curricular optimal transport on the recent state-of-the-art cross-lingual image-text method CCLM [19] during finetuning. As shown in Tab. II, applying our proposed dual-view curricular optimal transport for noise correspondence learning consistently brings improvement in all languages. Note that CCLM does not consider the noise of training data. These results not only demonstrate the good expansibility of our method to cross-lingual image-text similarity learning, but also further verify the effectiveness of our noise correspondence learning.

Comparisons on Multi-MSCOCO. Tab. III reports the experimental results on Multi-MSCOCO. DCOT significantly outperforms large-scale pre-trained models that do not consider the noisy correspondence problem. Compared to NRCCR, DCOT still has a huge advantage, with a 1.7% and 1.6% improvement in terms of sumR. Notably, compared to German and French in Multi30K, Chinese and Japanese in Multi-MSCOCO exhibit significant structural differences from English, making them more susceptible to noise during the translation process. Thus, Multi-MSCOCO is more challenging than Multi30K. The better performance of DCOT on Multi-MSCOCO than on Multi30K further demonstrates its superior robustness against noise. Besides, applying our method to CCLM also achieves a significant performance gain, showing that our method is compatible with popular pre-training models.

2) Cross-lingual Video-Text Retrieval: For cross-lingual video-text retrieval, we compare our model with two SOTA methods, MMP [15] and NRCCR [18]. Among them, MMP is pre-trained on Multi-HowTo100M (the multi-lingual version of HowTo100M [44]), and NRCCR is the robust learning
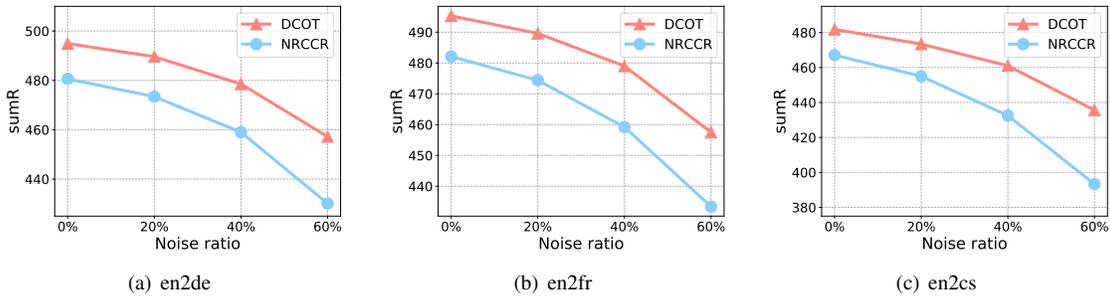
Fig. 7. Performance comparison with different noise ratios on Multi30K. The noise is injected by switching the correspondence of (V, T) pairs artificially, where "0%" means no noise is added artificially.

TABLE IV
PERFORMANCE COMPARISON OF CROSS-LINGUAL VIDEO-TEXT RETRIEVAL ON VATEX (THE SOURCE LANGUAGE IS ENGLISH AND THE TARGET LANGUAGE IS CHINESE). SYMBOL ASTERISK (*) INDICATES THE MODEL IS PRE-TRAINED ON MULTI-HOWTO100M [15].

Method	Text $\rightarrow$ Video			Video $\rightarrow$ Text			sumR
	R@1	R@5	R@10	R@1	R@5	R@1	
MMP [15]	23.9	55.1	67.8	-	-	-	-
MMP [15] *	29.7	63.2	75.5	-	-	-	-
NRCCR [18]	30.4	65.0	75.1	40.6	72.7	80.9	364.7
DCOT	**31.4**	**66.3**	**76.8**	**46.0**	**76.3**	**84.8**	**381.8**

TABLE V
ZERO-SHOT RESULTS ON MULTI30K. "-MT" INDICATES THE MULTI-LINGUAL VERSION. NOTE THAT CC3M AND CC3M-MT BOTH CONTAIN 3,346,732 IMAGE-TEXT PAIRS, WHILE MULTI-MSCOCO ONLY CONTAINS 616,435 IMAGE-TEXT PAIRS. THE RESULTS SUGGEST THAT NOISE-ROBUST LEARNING CAN ALLEVIATE THE DEPENDENCE ON LARGE-SCALE TRAINING DATA.

Method	Training Data	en2de	en2fr	en2cs
M^{3}P [17]	CC3M + Wikipedia	220.8	162.6	122.4
UC^2 [14]	CC3M-MT	375.0	362.4	330.6
CCLM [19]	CC3M-MT	409.5	384.4	375.3
NRCCR [18]	Multi-MSCOCO	448.7	433.8	411.2
DCOT	Multi-MSCOCO	**458.9**	**445.3**	**424.2**

method against noise introduced by MT. As shown in Tab. IV, DCOT demonstrates superior performance compared to large-scale pre-trained model MMP, which verifies the benefit of mitigating the noisy correspondence problem. Compared to the best baseline NRCCR, DCOT outperforms it by 4.7% in terms of sumR. From the results, one could see that our DCOT achieves excellent results, with the best results for cross-lingual video-text retrieval.

D. Generalization Analysis on Out-of-domain Data

In this section, we evaluate the generalization capability on out-of-domain data of our proposed method. In Tab. V, we report results on cross-lingual image-text retrieval under the zero-shot setting, where training data and testing data are from different domains. Compared with large-scale pre-trained model M^{3}P and CCLM, DCOT and NRCCR obtain superior results with fewer training data. Recall that both DCOT and NRCCR are trained in noise-robust manners.
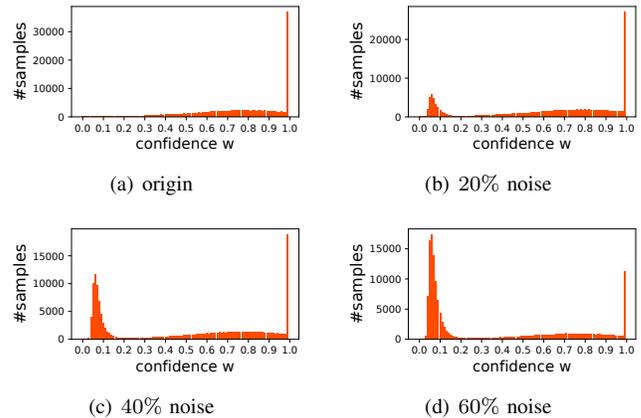

Fig. 8. Distribution of estimated confidence w under different noise ratios on French of Multi30K.

The results suggest that noise-robust learning can alleviate the dependence on large-scale training data and verifies the essential of noise-robust learning. In addition, with the same training data, our method consistently outperforms the noise-robust learning method NRCCR by a significant margin. This result verifies our proposed method DCOT has better cross-lingual transfer ability under the noisy scenario.

E. Robustness Analysis

To investigate the robustness of our proposed model, we conduct experiments with four different noise ratios on Multi30K. We compare our method with NRCCR which is the only noise-robust learning method for CCR. Specifically, we corrupt the training data by switching the correspondence of (V, T) pairs of some random instances based on a noise rate parameter. The higher the noise rates, the more serious the noisy correspondence problems become. The performance curves on two languages with the artificial noise ratios $[0, 0.2, 0.4, 0.6]$ are shown in Fig. 7. Our DCOT consistently performs better than NRCCR, and the performance gap between DCOT and NRCCR becomes larger as the noise rate increases. The results clearly show that DCOT performs more stability than NRCCR.

Video				
English	a man playing a guitar and a woman playing a violin are playing music on a sidewalk	a person diving and looking at the water in the ocean during the day	a man is surrounded by a court on an outdoor basketball court and dunking a basketball	a person is demonstrating how to thread a wide eyed needle with yarn
Translation (Chinese)	一个男人弹吉他和拉小提琴的女人正在人行道上播放音乐（a man playing a guitar and a woman playing a violin are playing music on a sidewalk）	在白天潜水和看海洋中的水 (diving and looking at the water in the ocean during the day)	一个男人被室外篮球场上的法院包围，扣篮篮球 (a man is surrounded by a court of justice on an outdoor basketball court, dunking basketball)	一个人正在展示如何用纱线向睁大眼睛针线 (a man is showing how to sew with yarn to open eyes)
Result	Cosine-sim: (0.26) DCOT: (0.90)	Cosine-sim: (0.29) DCOT: (0.88)	Cosine-sim: (0.49) DCOT: (0.39)	Cosine-sim: (0.54) DCOT: (0.36)

(a)

Video				
English	a young child sits on a floor and carefully folds a piece of clothing	a young child is behind a shopping cart and is pushing it	a young child is in a kitchen and pushes a stool under a table	a boy is at his track meet participating in the triple jump
Translation (Chinese)	一个幼儿坐在地板上，小心翼翼地折叠一件衣服 (a young child sits on a floor and carefully folds a piece of clothing)	一个幼儿在购物车后面，正在推动它 (a young child is behind a shopping cart, pushing it)	一个幼儿在厨房里，在桌子下推凳 (a young child is in a kitchen, he is under the table and pushes a stool.	一个男孩在他的轨道上迎接三跳 (a boy greets three jumps on his rail)
Result	Cosine-sim: (0.30) DCOT: (0.89)	Cosine-sim: (0.34) DCOT: (0.80)	Cosine-sim: (0.38) DCOT: (0.66)	Cosine-sim: (0.43) DCOT: (0.39)

(b)

Video				
English	a young boy holds a chocolate donut before taking a bite out of it	a person receives a neck massage while they are lying on their stomach	as the vehicle is up on a jack somebody is inspecting one of the tires and they see some damage	a boy in a black shirt has his arms raised and then he lowers them
Translation (Chinese)	一个年轻的男孩拿着巧克力甜甜圈，然后咬一口 (a young boy holds a chocolate donut before taking a bite out of it)	一个人在肚子上躺着颈部按摩 (a man lying on a belly for a neck massage)	随着车辆上的车辆上，有人在检查一个轮胎，他们看到一些伤害 (with the vehicle on the vehicle, someone was checking a tire and they saw some injuries)	一件黑色衬衫的男孩有他的手臂，然后他降低了他们 (a boy with a black shirt has arms, then lowers them)
Result	Cosine-sim: (0.31) DCOT: (0.94)	Cosine-sim: (0.28) DCOT: (0.87)	Cosine-sim: (0.39) DCOT: (0.40)	Cosine-sim: (0.49) DCOT: (0.30)

(c)

Fig. 9. Qualitative results of noisy correspondence learning on VATEX. We compare our proposed DCOT to the cosine similarity that is a point-to-point correspondence calculation method without considering the context. Our DCOT with context modeling achieves more reasonable confidence estimation results.

F. Visualization Analysis

1) Confidence Visualization: To further investigate the effectiveness of confidence estimation in our method, we carry out experiments by visualizing the per-sample confidence distribution under the different noise ratios. Specifically, we corrupt the training data by randomly switching the correspondence of (V, T) pairs of some instances using a specific noise rate. As the noise rates increased, the noisy correspondence problems became more severe. From Fig. 8, we can observe that the number of low-confidence samples increases with the noise ratio, which verifies the estimated confidence can reflect the magnitude of the noise.

2) Retrieval Visualization: In Fig. 9, we present the qualitative results on VATEX. From the results, we could observe that our DCOT could provide a more reasonable confidence estimation between sample pairs than the point-to-point counterpart that directly utilizes the cosine similarity to measure the correspondence without considering the context. Take Fig. 9 (a) as an example. The translation of the first example is significantly more accurate than the last one, and our DCOT

accordingly gives the highest score to the first one. In contrast, the cosine similarity based counterpart outputs the lowest score for this example. Moreover, we also observe that the cosine similarity scores vary in a very small range, which hardly reflects the quality of video-target language caption pairs. For the second example in Fig. 9 (c), our method fails to estimate confidence accurately. We assume that this is because the confidence estimation of our approach is based on global features, which capture the semantic information of a neck massage. However, it still has limitations in fine-grained information confidence estimation (*e.g.,* lying on a belly).

V. CONCLUSION

In this paper, we focus on the noisy correspondence problem in CCR. We propose a novel method, called dual-view curricular optimal transport (DCOT), which formulates the noisy correspondence learning in CCR as an optimal transport problem. We estimate the confidence of the correlated sample pair from both the cross-lingual and cross-modal views and design a dual-view collaborative curriculum learning strategy

to model the transportation costs dynamically according to the learning state of the two views. Additionally, we adjust the contribution of each data pair based on the estimated confidence to avoid network overfitting to the noisy sample pairs. Extensive experiments demonstrate the robustness and effectiveness of our method against noise introduced by MT. In future work, we plan to conduct an in-depth study of the impact of different levels of noise introduced by MT and explore how to adaptively adjust the curriculum schedule for different languages.

REFERENCES

[1] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once-multi-modal fusion transformer for video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 020–20 029.

[2] T. Han, W. Xie, and A. Zisserman, "Temporal alignment networks for long-term video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2906–2916.

[3] Y. Ge, Y. Ge, X. Liu, D. Li, Y. Shan, X. Qie, and P. Luo, "Bridging video-text retrieval with multiple choice questions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 167–16 176.

[4] M. Qi, J. Qin, Y. Yang, Y. Wang, and J. Luo, "Semantics-aware spatial-temporal binaries for cross-modal video retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 2989–3004, 2021.

[5] J. Dong, Y. Wang, X. Chen, X. Qu, X. Li, Y. He, and X. Wang, "Reading-strategy inspired visual representation learning for text-to-video retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[6] J. Li, L. Liu, L. Niu, and L. Zhang, "Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 9193–9207, 2021.

[7] Y. Zhang, W. Zhou, M. Wang, Q. Tian, and H. Li, "Deep relation embedding for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 617–627, 2020.

[8] L. Zhang and X. Wu, "Latent space semantic supervision based on knowledge distillation for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 31, pp. 7154–7164, 2022.

[9] J. Qin, L. Fei, Z. Zhang, J. Wen, Y. Xu, and D. Zhang, "Joint specifics and consistency hash learning for large-scale cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 31, pp. 5343–5358, 2022.

[10] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, and M. Wang, "Dual encoding for video retrieval by text," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4065–4080, 2022.

[11] P. Aggarwal and A. Kale, "Towards zero-shot cross-lingual image retrieval," *arXiv preprint arXiv:2012.05107*, 2020.

[12] J. Lei, T. L. Berg, and M. Bansal, "mtvr: Multilingual moment retrieval in videos," *arXiv preprint arXiv:2108.00061*, 2021.

[13] M. Portaz, H. Randrianarivo, A. Nivaggioli, E. Maudet, C. Servan, and S. Peyronnet, "Image search using multilingual texts: a cross-modal learning approach between image and text," *arXiv preprint arXiv:1903.11299*, 2019.

[14] M. Zhou, L. Zhou, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, "Uc2: Universal cross-lingual cross-modal vision-and-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4155–4165.

[15] P.-Y. Huang, M. Patrick, J. Hu, G. Neubig, F. Metze, and A. G. Hauptmann, "Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2443–2459.

[16] H. Fei, T. Yu, and P. Li, "Cross-lingual cross-modal pretraining for multimodal retrieval," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 3644–3650.

[17] M. Ni, H. Huang, L. Su, E. Cui, T. Bharti, L. Wang, D. Zhang, and N. Duan, "M3p: Learning universal representations via multitask multilingual multimodal pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3977–3986.

[18] Y. Wang, J. Dong, T. Liang, M. Zhang, R. Cai, and X. Wang, "Cross-lingual cross-modal retrieval with noise-robust learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, p. 422–433.

[19] Y. Zeng, W. Zhou, A. Luo, and X. Zhang, "Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training," *arXiv preprint arXiv:2206.00621*, 2022.

[20] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[21] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[22] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 406–29 419, 2021.

[23] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 308–14 317.

[24] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[25] M.-T. Luong, H. Pham, and C. D. Manning, "Bilingual word representations with monolingual quality in mind," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 151–159.

[26] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung *et al.*, "Multilingual universal sentence encoder for semantic retrieval," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[27] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2630–2640.

[28] C. Tan, J. Xia, L. Wu, and S. Z. Li, "Co-learning: Learning from noisy labels with self-supervision," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1405–1413.

[29] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 726–13 735.

[30] Y. Kim, J. Yim, J. Yun, and J. Kim, "Nlnl: Negative learning for noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 101–110.

[31] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, "Learning cross-modal retrieval with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5403–5413.

[32] D. Fu, D. Chen, H. Yang, J. Bao, L. Yuan, L. Zhang, H. Li, F. Wen, and D. Chen, "Large-scale pre-training for person re-identification with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2476–2486.

[33] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[34] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*.   PMLR, 2017, pp. 233–242.

[35] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," in *Proceedings of the 5th Workshop on Vision and Language*, 2016, pp. 70–74.

[36] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[37] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[38] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4581–4591.

[39] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 638–10 647.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[41] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[44] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.